STATISTICAL METHODS IN GENOME-WIDE ASSOCIATION STUDIES RELATED TO THE SEARCH FOR GENE BY RADIATION INTERACTIONS

Dan Stram
Preventive Medicine
University of Southern California

Outline

- What is meant by an "interaction"
- Some general observations on searching for gene x environment (GxE) interactions
- Gene x Radiation (GxR) specific requirements
- Suggestions for future research

Modeling Basics

- Generalized linear modeling for genes and exposure

$$g(\Pr(D)) = \alpha + \beta X + \gamma G + \psi G \times X$$

α - "intercept"
β - main effect of exposure X
γ - main effect of gene G
ψ -- "interaction" of G and X
g(.) – the "link function"

- When g() is the identity function then this an additive risk model, if an exponential function, then a relative risk model,
- For case control studies g() is the logistic function and if disease is "rare" then this is also (almost) a relative risk model

Epicure can allow for useful re-parameterizations

- For example we can constrain the estimation to fit an "excess risk" model

$$g(\Pr(D)) = \alpha[1 + \beta X + \gamma G]$$

as our main effects model and include an interaction as

$$g(\Pr(D)) = \alpha[1 + \beta X(1 + \psi G) + \gamma G]$$

here G modifies the excess risk associated with X

Interpretation of "interaction"

- Interpreting ψ depends on the link function g()
- In additive modeling we would view a purely multiplicative relationship between risk and X and G is an interaction (super-additive)
- In multiplicative modeling a purely additive relationship between risk and X and G is also an interaction (sub-multiplicative)

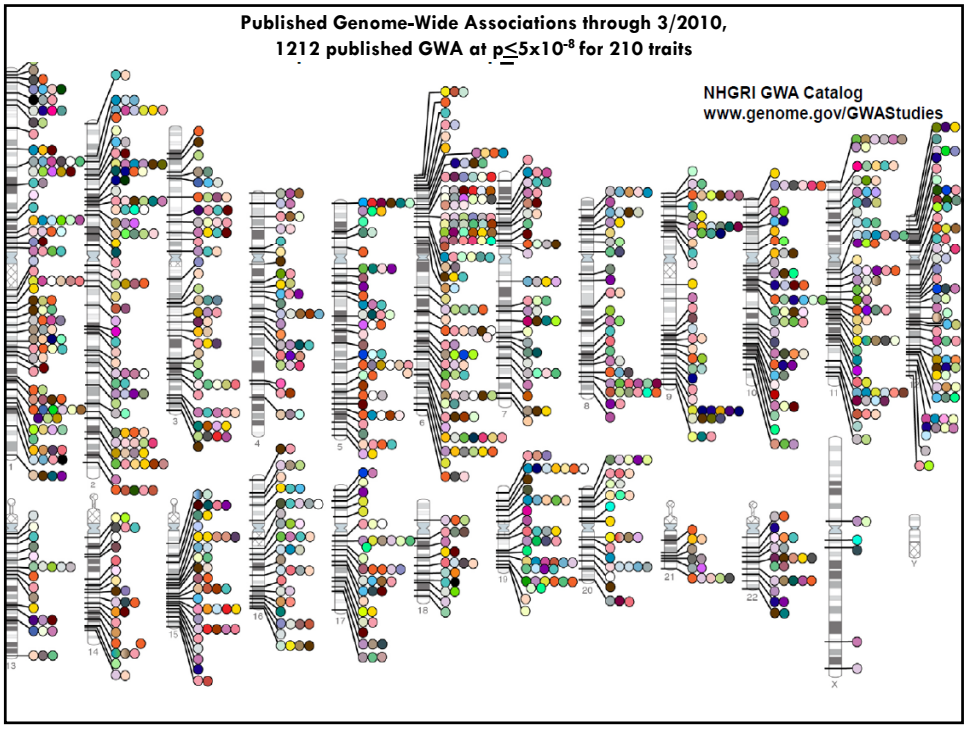
- How do we simplify things for the purpose of this talk?
  1. Assume that the risk factors X and G are relatively weak, so that we have little ability to distinguish additive from multiplicative effects
  2. Focus on "qualitative" interactions
     - If X has no effect on risk except in the presence of G then this is an interaction on any scale

GWAS studies

- Are dealing with relatively weak risk factors (for most cancers anyway)
- Are powered to be able to only detect the strongest (e.g. ~qualitative) interactions

GWAS studies

- Have yielded highly reproducible associations for many common alleles involved in risk of cancer and other diseases



Many main effects but precious few GxE interactions

- **Some interactions have been reported.**
  - A SNP in CFH and exposure to Chlamydia pneumoniae shows strong interactions in AMD progression (Baird et al, HMG 2008)
  - Smoking and SNPs in chromosome 15 for lung cancer appear to interact – but this may related to smoking behavior rather than susceptibility
  - There are hints of interactions with BMI for diabetes SNPs
- **Overall however there are few reliable interactions reported in the GWAS literature**

Why so few Gene by Environment Interactions in GWAS studies?

- The vast majority of GWAS findings in cancer have been relatively small in effect size (OR from 1.1 to 1.3 per copy) for variants with frequency above 5 percent.
- This may imply that GxE interactions for individual hits will also be small.

- Studies are not designed to detect modest interactions
  - Most are powered to detect risk alleles with consistent main effects
    - A risk allele that only raises risk in the presence of an exposure is less detectable in main effects analysis (unless we assume very large interactions)

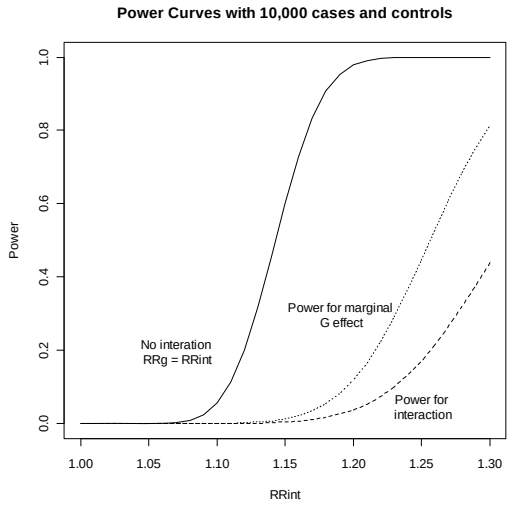
First main effects and then interactions

- Many studies look first for main effects and then explore interactions only for the top hits
  - Once a risk allele has been identified as a top hit then we have good power to define interactions – because multiple comparisons problem is strikingly reduced.
  - **On the other hand.** *SNPs with strong interactions may not be been identified in the initial scan*

Power Example

- Power considerations for a very common exposure (50 percent frequency) and common SNP (20 percent frequency)
  - Suppose that such a risk allele has no effect in the unexposed (RR_g=1) and a relative risk of RR_{int} >1 per copy in the exposed
  - Also assume the marginal affect of exposure is RR_e = 1.35
  - Consider the power of detecting either the marginal effect of this SNP or its interaction with exposure
    - After correcting for multiple comparisons (p<5e-8 nominal level using Bonferroni)

Power for GxE

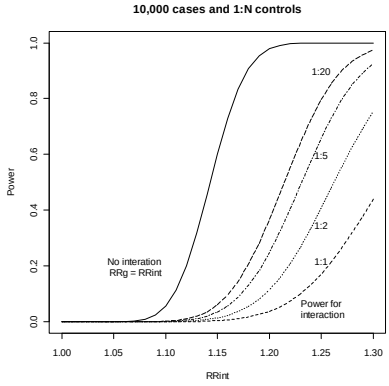


Implications

- The strategy of detecting interactions looking only among risk alleles with main effects can miss important variables involved in interactions (middle line on plot)
- The strategy of testing for interactions directly for all SNPs can be even weaker
  - This depends on the distribution of exposure: if an exposure is rare then it is more powerful to test for the interaction than the main effect – but only large interactions are detectable!

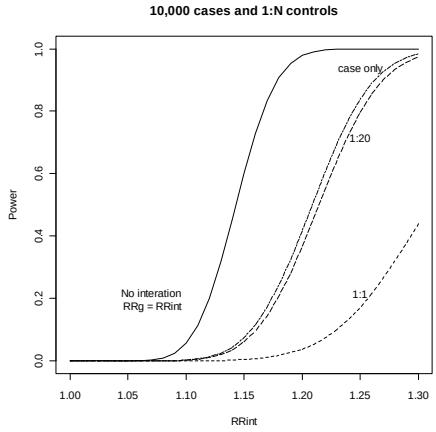
How to improve power

- Modify the design
  - In some studies controls are easier to recruit than cases (e.g. for a rare disease) what happens to power when we have more controls?



Case only analysis

- Analyze the data so that there are no controls



We have most power when there are zero controls. Why?

- Case-only analysis tests for association between risk alleles and the environment among cases
- The case-only analysis assumes that in the population as a whole the risk alleles are independent of environment
- Detecting an association between risk allele and environment in the cases is then tantamount to having detected a multiplicative interaction

However

- If there is an association between G and E in the population the case-only analysis will
  - Have greatly increased type I error if the risk alleles are positively associated with the environmental risk factor
  - Will have greatly reduced power if the risk alleles are negatively associated with E

Other approaches to increase power for interactions

- Counter-matching (Langholz and his co-workers)
  - Requiring that each risk set contains both exposed and unexposed cases improves power for detecting interactions
  - This ensures that fewer risk sets are "uninformative" for interactions
  - Has the effect similar to adding one or two controls for the interaction comparison – without extra genotyping!
  - No requirement for G x E independence
  - However it weakens the ability of the study to detect the main effect of risk alleles – unless we assume that risk alleles are independent of exposure

What issues are important in radiation epidemiology?

- Three most important issues seem to me to be
  1. Characterizing the synergy between already identified GWAS hits and radiation
  2. Identifying new risk alleles that only become important when radiation is present
  3. Translating confirmed associations into biological mechanism

Synergy between all the GWAS hits and Radiation

- Are people who are susceptible because of genetics also more susceptible to the effects of radiation overall?
- Simple weighted risk allele sums seem to do a good job in summarizing the joint effects of carrying the risk variants

$$\text{wsum}=(b_1G_1 + b_2G_2 + \ldots + b_nG_n)$$

  - For example a meta-analysis of height involving 134,000 individuals found 180 associated variants
  - No gene x gene interactions were noted!
- Such a score can be quite powerful predictor of risk even just using only the current GWAS hits
- We will have much more power for examining interactions with the risk score than with individual markers.

Synergy between GWAS hits and Radiation

- Here our interest is not restricted to detecting multiplicative interactions
  - Suppose that risk alleles G and radiation R follow the strictly log additive model for both G and R

$$\Pr(D=1) = \text{background} \times \exp(\text{wsum}(G) + \beta_r R)$$

If the total sum of effects of G is large then most radiation induced cancers will tend to be seen in the genetically sensitive even without a multiplicative interaction.

ERR models

- The same holds true for the more standard ERR models

$$\Pr(D=1) = \text{bkg}(\text{wsum}) \times (1 + b_r R)$$

- In this model as well more of the excess cancers due to radiation will again be seen among the genetically susceptible
- Thus for GxR studies it will be just as important to evaluate the ERR model as to identify individual risk alleles that interact with radiation

Modifiers in the ERR model

- We can treat a single G as a modifier of the effect of radiation by fitting

$$\Pr(D=1) = \text{bkg} \times \{1 + b_r \, \text{R} \exp(b_{gr} \times G)\}$$

- However again we may be more interested in

$$\Pr(D=1) = \text{bkg} \times \{1 + b_r \, \text{R} \exp[b_{gr} \times \text{wsum(G)}]\}$$

Pathway analysis

- As we learn more about the biology of the GWAS hits we should be able to group these into specific pathways and test whether the sum total of groups of genes modify the effect of radiation as in

$$\Pr(D=1) = \text{bkg} \times \{1 + b_r \, \text{R} \exp(b_{gr} \times \text{wsum}(\text{P}_1) + b_{gr2} \times \text{wsum}(\text{P}_2) + ...+\}$$

- Where wsum($\mathbf{P_1}$) is the weighed sum of risk alleles in pathway 1, etc. This allows the effect of radiation to be modified by the specific effects of risk alleles in homogeneous pathways

Case-only analysis reconsidered for specific radiation studies

- Consider causes of failure of G x R independence in the (appropriate) general population
  - In A-bomb study
    - Hard to think of how a gene would cause radiation exposure in this setting
    - . However there are legitimate questions about hidden population stratification
      - Urban rural differences relevant to some control groups
      - Many of the people living in the Hijiyama shadow were from an isolated sub-caste and may have some small genetic differences between other (more) exposed groups
    - In general however the case only test seems very "safe" in this study

Specific studies

- WECARE study (cases of contralateral breast cancer exposed or unexposed to radiation at time of first diagnosis)
  - The population in which we require G x R independence now is women with breast cancer
    - it is possible that some alleles determine tumor characteristics leading to decisions about radiation usage
      - So far studies have found limited relationship between SNPs and tumor characteristics but these studies are still in their infancy
      - The main tumor characteristics associated with radiation exposure would seem to be stage and size which are measured in the study and can be controlled for
    - Decisions to use radiation could possibly be related to genetically-related host characteristics (national origin, body shape or size)
      - These seem very remote

Suggestions for future research

- Joint modeling of known risk alleles
  - A primary question is whether radiation sensitivity varies by background genetic susceptibility
  - Using (weighted) sums of known risk alleles as a single genetic risk variable and looking for heterogeneity in radiation response according to degrees of "genetic" risk should be a key issue in future studies.
  - If we see no evidence for interaction (i.e. non-multiplicative) this is just as important (in a well powered study) as a finding of interactions would be.
  - Risk score analysis has considerably more power than does analysis of individual SNPs (each of which probably has modest interactions with radiation).

Acknowledgements

- USC Genetic Epidemiology and Statistical Genetics
  - Chris Haiman, Duncan Thomas, Brian Henderson, Gary Chen, David Conti, James Gauderman, etc. etc.